Calibrated breast density methods for full field digital mammography: A system for serial quality control and inter-system generalization

B. Lu, A. M. Smallwood, and T. A. Sellers
Department of Cancer Epidemiology, Division of Population Science, H. Lee Moffitt Cancer Center, Tampa, Florida 33612

J. S. Drukteinis
Department of Diagnostic Imaging, H. Lee Moffitt Cancer Center, Tampa, Florida 33612

J. J. Heine[a]
Department of Cancer Imaging and Metabolism, Division of Basic Science, H. Lee Moffitt Cancer Center, Tampa, Florida 33612

E. E. E. Fowler[a]
Department of Cancer Epidemiology, Division of Population Science, H. Lee Moffitt Cancer Center, Tampa, Florida 33612

Purpose: The authors are developing a system for calibrated breast density measurements using full field digital mammography (FFDM). Breast tissue equivalent (BTE) phantom images are used to establish baseline (BL) calibration curves at time zero. For a given FFDM unit, the full BL dataset is comprised of approximately 160 phantom images, acquired prior to calibrating prospective patient mammograms. BL curves are monitored serially to ensure they produce accurate calibration and require updating when calibration accuracy degrades beyond an acceptable tolerance, rather than acquiring full BL datasets repeatedly. BL updating is a special case of generalizing calibration datasets across FFDM units, referred to as cross-calibration. Serial monitoring, BL updating, and cross-calibration techniques were developed and evaluated.
Methods: BL curves were established for three Hologic Selenia FFDM units at time zero. In addition, one set of serial phantom images, comprised of equal proportions of adipose and fibroglandular BTE materials (50/50 compositions) of a fixed height, was acquired biweekly and monitored with the cumulative sum (Cusum) technique. These 50/50 composition images were used to update the BL curves when the calibration accuracy degraded beyond a preset tolerance of ±4 standardized units. A second set of serial images, comprised of a wide-range of BTE compositions, was acquired biweekly to evaluate serial monitoring, BL updating, and cross-calibration techniques.
Results: Calibration accuracy can degrade serially and is a function of acquisition technique and phantom height. The authors demonstrated that all heights could be monitored simultaneously while acquiring images of a 50/50 phantom with a fixed height for each acquisition technique biweekly, translating into approximately 16 image acquisitions biweekly per FFDM unit. The same serial images are sufficient for serial monitoring, BL updating, and cross-calibration. Serial calibration accuracy was maintained within ±4 standardized unit variation from the ideal when applying BL updating. BL updating is a special case of cross-calibration; the BL dataset of unit 1 can be converted to the BL dataset for another similar unit (i.e., unit 2) at any given time point using the 16 serial monitoring 50/50 phantom images of unit 2 (or vice versa) acquired near this time point while maintaining the ±4 standardized unit tolerance.
Conclusions: A methodology for monitoring and maintaining serial calibration accuracy for breast density measurements was evaluated. Calibration datasets for a given unit can be translated forward in time with minimal phantom imaging effort. Similarly, cross-calibration is a method for generalizing calibration datasets across similar units without additional phantom imaging. This methodology will require further evaluation with mammograms for complete validation. *© 2015 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution 3.0 Unported License.* [http://dx.doi.org/10.1118/1.4903299]

1. INTRODUCTION

Breast density is an important breast cancer risk factor most often estimated from mammograms. There are various methods under investigation for measuring breast density[1,2] including those that operate on the image data directly or incorporate calibration. Calibration, or standardization, methods are designed to account for image acquisition technique differences and are more recent developments in mammography.[3–10] Commercial products using standardization are also available to estimate breast density.[11,12] Similarities in the various calibration approaches and related breast density measures were discussed in our previous work.[3,13]

When developing a standardized inter-mammogram data (i.e., pixel) representation, calibration data are normally acquired or developed from the respective mammography unit(s). In general, there are different ways to collect these data. In one approach, for example, data can be collected simultaneously at the time of the mammogram acquisition by installing a calibration device in the mammography system.[8,9] We refer to this as an internal method. Alternatively, calibration data can be acquired without a hard-system interface, which we refer to as an external method.[7] There are both benefits and drawbacks with either approach. The internal approach collects data in real-time, mitigating serial drift influences. External approaches that use phantom imaging may experience serial calibration accuracy degradation if the mammography unit experiences drift relative to a calibration dataset collected prior to patient imaging. The external approach may have more latitude in the calibration datatype because the data collection is not visible to the clinical operation. The most appropriate method for standardization is still under investigation. Because we are using an external approach with phantom imaging,[3,4,14,15] it is necessary to establish a serial monitoring component in the calibration system.

We are operationalizing the cumulative sum (Cusum) technique[16] for monitoring serial calibration accuracy.[17] Cumulative sum has its origins in industrial process quality control[18–20] and is suited for detecting sustained drift relative to a reference. Since its original development, Cusum has been used in various arenas including medical outcomes evaluation and public health surveillance. Such applications include monitoring performance of clinicians in various specialties,[21–30] patient survival following transplant procedures,[31] screening process quality,[32] infectious disease outbreaks,[33–36] cancer incidence,[37] air pollution levels,[38] and occupational safety events.[39]

Our calibration methodology was initially developed using indirect x-ray conversion full field digital mammography (FFDM) technology.[3,4,15,17] We are currently modifying these methods for use with direct x-ray conversion FFDM.[14] In this paper, we use phantom imaging to evaluate the serial stability of three direct x-ray conversion FFDM units relative to calibration accuracy, assess methods for maintaining accuracy using the Cusum technique, and evaluate a basis for applying calibration data collected from one unit to another similar unit for generalization purposes.

2. MATERIALS AND METHODS

2.A. Background

Our calibration approach relies on acquiring a baseline (BL) calibration dataset at time zero (BL_0) for each FFDM unit. Each calibration dataset requires a considerable amount of breast tissue equivalent (BTE) phantom imaging (about 160 acquisitions), making it impractical to collect full calibration datasets repeatedly in time to maintain prospective accuracy. Instead, we monitor each unit serially to assess the applicability of its BL_0 with a minimal amount of serial phantom imaging, requiring approximately 20 minutes biweekly per mammography unit. When sustained variation beyond a preset level is detected at some time point after establishing BL_0 in a given unit, its calibration dataset can be updated (by hypothesis) using the serial phantom images, essentially bringing its BL_0 dataset forward in time. The basis for this updating mechanism is a special case of more general theory that shows how a calibration dataset acquired with one FFDM unit can be converted to another similarly manufactured unit using a minimal amount of phantom imaging, referred to as cross-calibration. Both the special case of serial calibration dataset updating for a given unit and the more general theory of cross-calibration are presented and evaluated in this paper.

2.B. Mammography units

Phantom images were acquired from three Hologic Selenia direct x-ray conversion FFDM units. These units are used for breast screening and diagnostic purposes at Moffitt Cancer Center. The Selenia detector has 70 micron pitch (pixel resolution). These units produce raw images, stored with 14 bit per pixel dynamic range, and clinical display images, stored with 12 bit per pixel dynamic range. Raw images were used for our work. For screening mammography, these systems primarily use two detector field of views (FOVs) determined by the choice of compression paddle. We examined data acquired with the large FOV (24×29 cm or 3328×4096 pixels) in this paper, because the choice of FOV does not affect the calibration accuracy.[14] Two units (H_1 and H_2) have tungsten/rhodium (W/Rh) and tungsten/silver (W/Ag) target/filter combinations. The third unit (H_3) has molybdenum/molybdenum (Mo/Mo) and molybdenum/rhodium (Mo/Rh) combinations.

2.C. Materials

Images of BTE phantom (CIRS, Norfolk, VA) were used for this work. Our phantom sets consist of 100% adipose and 100% fibroglandular (glandular) BTE materials with 1 mm, 2 mm, 1 cm, or 2 cm thickness (i.e., precise heights). The area dimensions are 18 cm $\times$ 24 cm for the large set and 12.5 cm $\times$ 10 cm for the small set. For a given set, composite compositions were constructed by stacking combinations of these homogeneous BTE materials. For example, we let $h = h_g + h_a$, where h is the total height in centimeters of a given stacked phantom arrangement (i.e., height equates with the compressed breast thickness) measured in centimeters, h_g is the height of the glandular

component, and h_a is the height of the adipose component. The theoretical (ideal) percent glandular (PG) standardized value of a given composite is then given by $100 \times h_g/h$. Composite phantoms are referenced by the glandular percentage/adipose percentage. For example, 40/60 references a composite phantom comprised of 40% glandular and 60% adipose BTE materials for a given total height and is referred to as 40 PG. More generally, a composite composition is referenced as w/z below, where $w + z = 100$, w gives the PG designation, and z is the percentage of adipose tissue by height.

2.D. Calibration

The calibration methods were described previously[3,4,14] and are outlined here. Initial BL_0 datasets are acquired prior to calibrating patient mammograms and are used to establish calibration curves. These curves are functions of the acquisition techniques shown in Table I for each unit. There are 16 acquisition techniques sampled for the H_1 and H_2 units (similar units) and 15 acquisition techniques for the H_3 unit. To limit the amount of calibration data collection, we only sampled the range of compressed breast thicknesses routinely observed in practice. For thicknesses beyond this range, we used extrapolation based on the linear regression model described previously[4] considering sampled heights ≥ 4 cm in the model. For each acquisition technique, there are two calibration curves corresponding to the 100% adipose and 100% glandular BTE phantom images. These curves and calibration are developed in the logarithm of the relative exposure (LRE) representation as a function of compressed breast thickness (i.e., phantom height). When setting $x = $ mAs (i.e., the acquisition mAs) and

letting the value of a given pixel or the average value of a group of pixels = pv, the LRE representation is given by $\ln(pv/x)$.

Shorthand notation is used for the developments below. For time points other than time zero, we replace the zero subscript with the index n, unless noted, where n is a running index defining the number of serial samples taken since time zero that can be converted to the total number of days since time zero. Adipose and glandular calibration curves for a given acquisition technique corresponding to BL_n are referred to as A_n and G_n, respectively. When height considerations are required, we reference these as $A_{n,h}$ and $G_{n,h}$, where h is the total phantom height in centimeters. Given a LRE corresponding to an arbitrary w/z composite phantom [i.e., LRE(w)], we express the calibration application symbolically as CAL[LRE(w)] = w PG units paralleling the description in Sec. 2.C.

2.E. Serial imaging

In addition to the BL datasets, two series of phantom images were acquired from each FFDM unit defined as Timeline 1 and Timeline 2. Timeline 1 was comprised of 50/50 composition phantoms (large set) acquired biweekly using the acquisition techniques shown in Table I with $h = 4$ cm. Due to study timing, the starting date and the duration of the Timeline 1 imaging vary across the units. Imaging was performed from July 2012 to present for H_1 and H_2 and from March 2013 to present for H_3. The phantom position and region of interest (ROI) used for the Timeline 1 acquisitions, the BL_0 curve generation, and analyses are shown in Fig. 1 (left). Timeline 1 images were used for multiple purposes: (i) serial monitoring with Cusum

TABLE I. Acquisition techniques. This table gives the acquisition techniques and phantom height ranges for the BL calibration datasets for each FFDM unit. H_1 and H_2 have W/Rh and W/Ag target/filter combinations, whereas H_3 has Mo/Mo and Mo/Rh combinations. The BL datasets for H_1 and H_2 were acquired in July 2012 and in March 2013 for H_3 unit. As indicated in the footnotes, additional techniques were added to these BL datasets afterward.

FFDM unit	Target/filter	KV	Height range (cm)	FFDM unit	Target/filter	KV	Height range (cm)
H_1/H_2	W/Rh	25[a]	n/a	H_3	Mo/Mo	25	2–6
		26	2–6			26	2–6
		27	2–6			27	2–6
		28	2–7			28	2–7
		29	3–7			29	3–7
		30	3–7			30	3–7
		31[a]	n/a			31	3–7
		32[a]	n/a			32	n/a
	W/Ag	26	n/a		Mo/Rh	26[b]	3–7
		27	3–7			27	3–7
		28	3–7			28	3–7
		29	3–7			29	3–7
		30	3–7			30	3–7
		31	3–7			31	3–7
		32	3–7			32	3–7
		33[a]	n/a			33	3–7
		34[a]	n/a			34	n/a

[a]Additional acquisition techniques that were added during September 10–11, 2013.
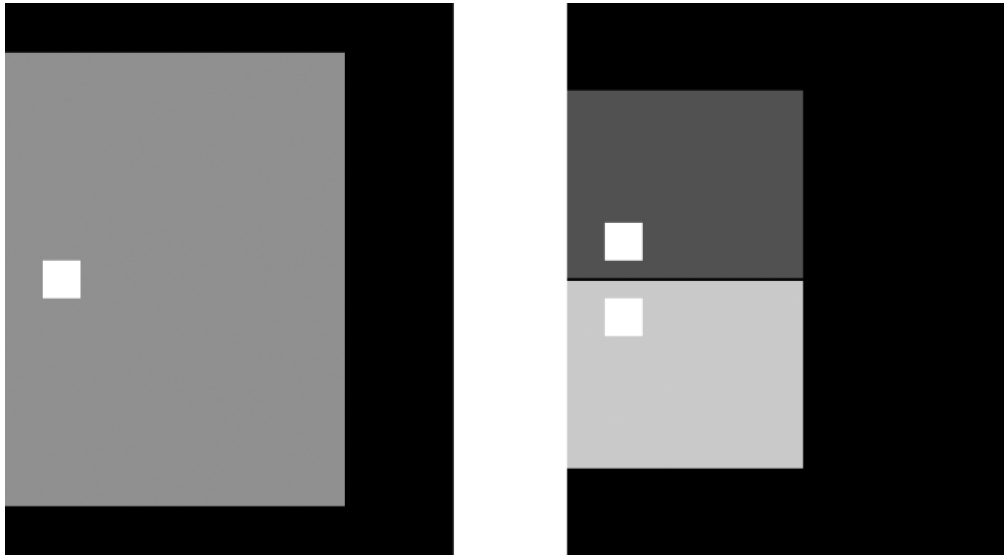[b]The acquisition technique that was added on April 01, 2013.

FIG. 1. Phantom positions and ROIs. This shows the phantom locations illustrated by various gray shaded rectangles relative to the detector FOV, shown as the larger black rectangle, and the respective ROIs, illustrated with white-squares: (i) a large-area phantom placement used for Timeline 1 imaging (left layout); and (ii) the position of two small-area phantoms placed side by side (right layout) used for the Timeline 2 acquisitions.

as described below; (ii) BL_0 updating when required to correct out of control (OC) behavior; and (iii) cross-calibration accuracy evaluation. For Timeline 2, four different compositions were imaged biweekly for each acquisition technique with $h = 5$ cm. This height was used because it was within the expected compressed breast thickness range and added variability to the serial accuracy evaluation. To minimize the amount of imaging effort while acquiring range of composite phantoms for Timeline 2, two different compositions were imaged simultaneously using the small-area phantoms for each acquisition technique, as shown in Fig. 1 (right): (1) 40/60 and 60/40 compositions, and (2) 20/80 and 80/20 compositions. Timeline 2 images were used to evaluate the serial calibration accuracy derived from either BL_0 or the updated BL_n for each unit and to evaluate the cross-calibration. Timeline 2 imaging was initiated at a later date relative to Timeline 1 (March 2013—present for all three units). Timeline 2 images offer a means for independent evaluation because they were not used in the BL updating analysis.

2.F. Cumulative sum monitoring

The sequential Cusum technique was used to construct the decision interval (DI for shorthand) Cusum monitoring.[16] We applied the DI method described in our previous work[17] with modifications to monitor in control (IC) behavior and detect sustained variation using Timeline 1 images. The DI detects upward and downward drift (separately) from a defined reference relative to a preset tolerance using two respective time series. The LRE of a 50/50 composition taken at or near BL_0 (time zero) defines the initial reference for each acquisition technique, $m_0 = LRE_0 (50)$. The LRE at n, expressed as $m_n = LRE_n (50)$, is used to define the Cusum variable

$$U_n = \frac{m_n - m_0}{|m_0|}, \tag{1}$$

for $n > 0$ with $U_0 = 0$. The above relationship sets the monitoring to a relative shift from m_0.[17] The standard sequential Cusum is given by

$$S_n = \sum_{i=1}^{n} S_{n-1} + U_n. \tag{2}$$

Briefly, Eq. (2) is the running total of the deviation from the reference for the first n samples[16] in normalized form. For random variation, the plot of S_n versus n is random (i.e., S_n is a zero mean random variable). When the drift is sustained in a specific direction, S_n will veer off in the same direction. Using Eq. (1), the respective DI forms for detecting upward and downward shifts are given by

$$S_n^+ = \max(0, S_{n-1}^+ + U_n - k) \tag{3}$$

and

$$S_n^- = \min(0, S_{n-1}^- + U_n + k), \tag{4}$$

where $S_0^- = S_0^+ = 0$, k is the chart constant, and the normalized LRE shift tolerance defined below. When the deviation is less than this tolerance, the DI terms return zero indicating IC behavior. Otherwise, the DI triggers and returns a value other than zero from one of its terms depending on the drift direction. We define OC behavior as two consecutive upward or downward LRE shifts exceeding the tolerance (i.e., two consecutive DI triggers in the same direction). When OC behavior is detected, calibration curves and the Cusum are adjusted accordingly to bring the system into tolerance. For example, when this event occurs at n, BL_0 is updated at $n-1$ giving BL_{n-1} (i.e., effectively creating a new BL_0) and the Cusum is reset: $m_0 = m_{n-1}$ and $S_{n-1} = S_0 = 0$. For each BL update, k is recalculated with the new m_0 to maintain the ± 4 PG tolerance after the reset. Figure 2 shows the schema for the monitoring and updating. Typically, Cusum applications use a detection threshold in conjunction with the chart
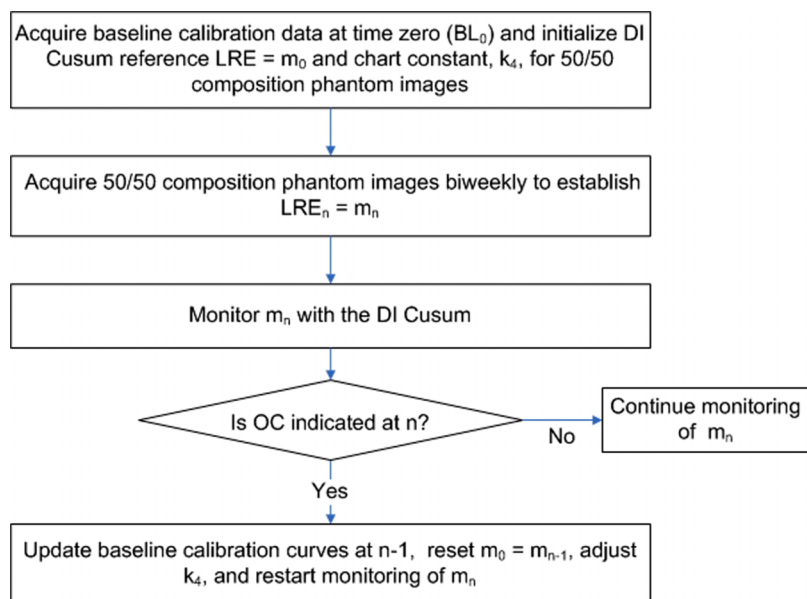
FIG. 2. Serial monitoring schema for Timeline 1. Out of control (OC) behavior is defined as two consecutive DI Cusum triggers, occurring at $n-1$ and n. The reset and updating take place retrospectively at $n-1$ relative to the OC event detection at n.

constant[16] to define OC behavior. In our application, we use the two consecutive DI trigger rules developed specifically for this application based on our choice of k, rather than the threshold approach.

The chart constant, k, plays a critical role in capturing OC behavior and is equivalent to the fractional shift from the initial LRE reference for a given acquisition technique (i.e., kV and target/filter) in our application. We used methods developed previously[17] to determine k with further modifications. For reference, there is a linear and symmetric relationship between the degree of deviation from the reference LRE and the resulting variation in PG (i.e., calibration error) for a fixed height. Additionally, the variation in the calibration accuracy is a function of the acquisition technique and height, but not of composition. For all acquisition techniques and heights, we required the serial calibration accuracy to be within a ±4 PG tolerance from the ideal value,[14] forming the basis for this user-imposed IC bound. The ±4 PG width defines the 95% confidence interval (CI) for the calibration accuracy. To achieve this uniform accuracy objective, look-up tables (LUTs) were developed numerically and modeled with polynomials to estimate the appropriate k because we only monitor $h = 4$ cm with actual serial imaging.

We outline the methods to estimate k below and provide the mathematical details in the Appendix. The 50/50 reference LREs are established from the BL_0 dataset using a linear combination. We numerically find the values of k for each acquisition technique that gives a ±4 PG shift in the calibration of the Timeline 1 50/50 composition images ($h = 4$ cm) giving k_4. A similar approach is then used to estimate k for the full range of heights (i.e., for $h \neq 4$ cm) using a normalization that references arbitrary $k - k_4$ resulting in height dependent LUTs. The k for arbitrary h is found by taking the appropriate point from the LUT and adding it to k_4 giving k_h. Thus 50/50 compositions for arbitrary h can be monitored using Timeline 1,

essentially creating virtual timelines. The justification for the simultaneous monitoring for all h based on one serial timeline (i.e., $h = 4$ cm) follows from the developments provided in Sec. 2.G. below and is illustrated in Sec. 3.

2.G. Baseline updating and cross-calibration theory

Timeline 1 images are used to update BL_0 (or more generally BL_n) when the DI detects OC behavior. The goal of updating is to reconstruct the BL_0 dataset at any time after its initial collection using one phantom image per acquisition technique taken near the time of reconstruction. We make these assumptions used in the updating development outlined below: (i) various difference equations (described below) derived from calibration curves at arbitrary time points are statistically time-invariant; and (ii) BL_n datasets within or across similar Hologic units vary by an offset (i.e., a constant shift). With these assumptions, the following hypotheses were evaluated: (i) the BL_0 dataset for a given acquisition technique and unit can be updated (i.e., reconstructed) using one 50/50 composite phantom image per acquisition technique taken at n giving BL_n. This essentially translates BL_0 in time to BL_n by making the appropriate adjustments and is a special application of a more general theory. (ii) For arbitrary n and m, the BL_n from H_1 can be converted to BL_m for H_2 (or vice versa) by using the other similar unit's 50/50 compositions taken at n or m, defining the more general theory of cross-calibration.

The BL updating (translation) and cross-calibration operations follow from the same development. The theoretical calibration curves for a given unit at n are related to a composite composition acquired at n (or near n) by a linear combination

$$\mathrm{LRE}(c \times 100)_{n,h} = [(1-c) \times A_{n,h} + (c) \times G_{n,h}], \tag{5}$$

where c is the [0,1] valued mixing coefficient with $c = 1/2$ for a 50/50 composition. We solve Eq. (5) for $h = 4$ cm with

$c = 1/2$ giving one calibration curve point at n,

$$G_{n,4} = 2 \times \mathrm{LRE}(50)_{n,4} - A_{n,4}. \tag{6}$$

We use a difference equation relating BL_0 with BL_n ($n > 0$) to define an increment

$$\Delta_h = A_{0,h} - G_{0,h} = A_{n,h} - G_{n,h}. \tag{7}$$

Equation (7) is the difference between a given unit's calibration curves as a function of h and is valid under our assumptions because BL_0 can be acquired at any time. We note, this difference is the contrast between the adipose and glandular pixel values for a given h in the LRE representation. Also by hypothesis, this difference is a characteristic quantity of similarly manufactured units (i.e., H_1 and H_2). Rearranging the above equations with substitution gives

$$A_{n,4} = \mathrm{LRE}(50)_{n,4} + \frac{1}{2}\Delta_4 \tag{8}$$

and

$$G_{n,4} = \mathrm{LRE}(50)_{n,4} - \frac{1}{2}\Delta_4. \tag{9}$$

Equations (8) and (9) specify one point from each theoretical calibration curve at n in terms of a difference from the BL_0 dataset and a measured 4 cm LRE for a 50/50 composition acquired at n. The next differences resemble discrete *derivatives* with respect to h expressed as

$$\mathrm{d}a_h = A_{0,h} - A_{0,h+1} = A_{n,h} - A_{n,h+1} \tag{10}$$

and

$$\mathrm{d}g_h = G_{0,h} - G_{0,h+1} = G_{n,h} - G_{n,h+1}. \tag{11}$$

Equations (10) and (11) provide the means to derive the calibration curves at n using $A_{n,4}$ and $G_{n,4}$ for a given unit or develop the other unit's BL_n (i.e., cross-calibration) by the choice of 50/50 composition used in this updating scheme (i.e., either acquired from H_1 or H_2). The general equation for the adipose calibration curve reconstruction at n for $h > 4$ cm is given by

$$A_{n,h+1} = A_{n,h} - \mathrm{d}a_h. \tag{12}$$

For example, the point for $h = 5$ cm is given by

$$A_{n,5} = A_{n,4} - \mathrm{d}a_4. \tag{13}$$

The general expression for $h \le 4$ cm is expressed as

$$A_{n,h} = A_{n,h+1} + \mathrm{d}a_h. \tag{14}$$

For example, the point for $h = 3$ cm is given by

$$A_{n,3} = A_{n,4} + \mathrm{d}a_3. \tag{15}$$

The other points are derived similarly, and the solution for $G_{n,h}$ follows the same form. The development outlined above shows how to reconstruct the entire BL calibration dataset with one image per acquisition technique, and its merits were evaluated with additional experiments described below. The assumptions used above for the difference relationships are supported by the linear form of the calibration curves.[4,14,15]

We used a 50/50 composition with $h = 4$ cm for this development, but these choices are not unique. The 50/50 composition gives even weight to the A and G curves (i.e., reducing bias because the mixing coefficients are equivalent $c = c_{\mathrm{glandular}} = 1 - c_{\mathrm{adipose}} = 1/2$), and 4 cm represents a central range for the expected compressed breast thickness. We note in practice, the BL curves are established (sampled) at integer height values. A cubic-spline interpolation is used to estimate intermediate points as required.[14]

Three additional experiments were designed to evaluate the BL update/translation and cross-calibration hypotheses: (i) to evaluate the BL translation within a given unit, we updated BL_n at each n for all $n \ge 1$ and calibrated the respective unit's Timeline 2 images acquired at each n and evaluated accuracies with respect to the tolerance (i.e., within-unit evaluation), referred to as continual BL translation; (ii) we performed cross-calibration at each $n \ge 0$ for H_1 and H_2 using their Timeline 2 images at n and made comparisons with the tolerance and the continual within-unit BL translation; and (iii) as an additional means for the cross-calibration comparison, calibration data acquired from H_1 was applied to the Timeline 2 images acquired with H_2 and vice versa at each n for all $n \ge 0$ without modification (i.e., switching BL_0 calibration datasets referenced as Cal-switch), in contrast with the cross-calibration evaluated above.

3. RESULTS

3.A. Decision interval cusum monitoring

To establish the serial monitoring, chart constants were estimated numerically as a function of acquisition technique and height for each unit. Figure 3 shows examples of the adjustment LUTs with the k_4 influence removed for the H_1 and H_3 units. The modified k was determined by adding the respective adjustment from the LUT to k_4. The adjusted k detects ± 4 PG shift in the calibration at a given height using the Timeline 1 series (i.e., 4 cm phantoms) as the reference (all acquisition techniques). Data for H_2 are not shown due to its similarity with H_1 data. The chart constant adjustments were modeled as a function of h with a fifth degree polynomial. The plots show the k adjustments and h increase in tandem indicating that the calibration accuracy has greater uncertainty as h decreases. This greater uncertainty is due to the separation between the adipose and glandular calibration curves, which decreases as h decreases. This effect is illustrated in Table II by showing: (i) various LRE shifts (absolute value) for a 50/50 composition (W/Ag at 27 kV) required at specific heights to cause a 4 PG calibration shift at each of the specified heights; and (ii) the calibration errors that this shift induces at the nonspecified heights. As a specific example, the italicized row shows that a 0.0383 LRE shift at 5 cm translates to these errors: (i) 4 PG shift at 5 cm as expected, (ii) 5.71 PG shift at 3 cm, and (iii) 3.27 PG shift at 7 cm. Similarly, these plots show that smaller variations in the LRE are required to induce a 4 PG shift as kV increases for a given height and target/filter combination, as the calibration curves also become less separated.
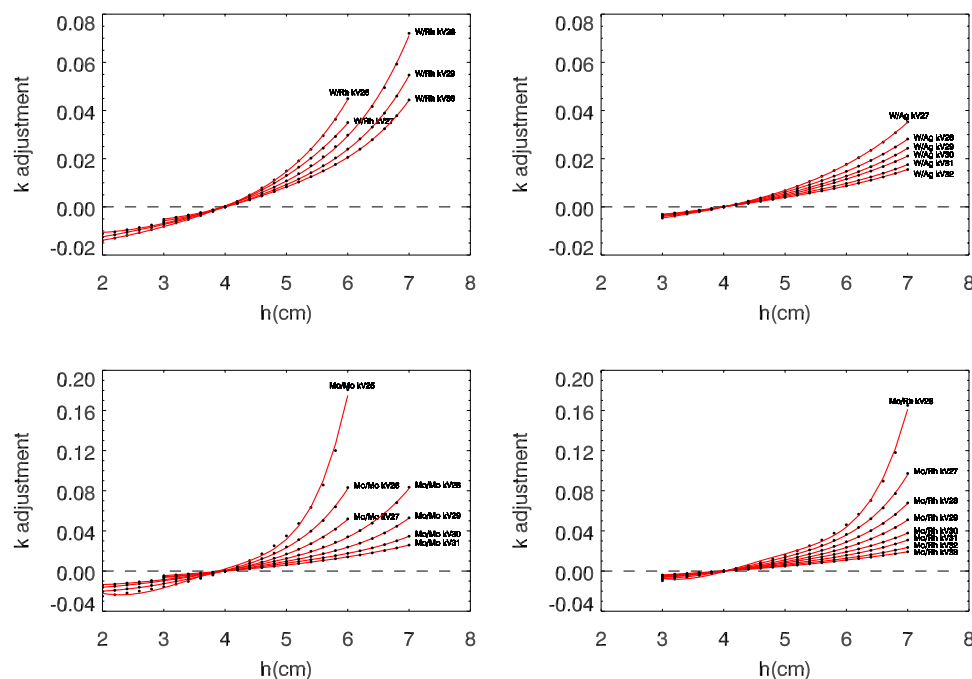
FIG. 3. Chart constant LUTs. This illustrates the chart constant LUTs for H_1 (top left and right) and H_3 (bottom left and right). Sample points (dots) were generated numerically using measured values at time zero and represent relative k adjustments with the influence of k_4 removed. These were derived from the baseline dataset for each target/filter and kV combination. The fitted fifth degree polynomials (solid curves) represent the relative adjustments for the chart constant as a function of height (h). The adjustment is determined from the appropriate curve at the respective time (i.e., for a given target/filter and kV) given $h \neq 4$ cm, which is then added to k_4 for the DI monitoring and is required to detect a ± 4 PG shift.

To further illustrate the importance of using the appropriate k and BL updating when required, the DI was first applied over the entire Timeline 1 interval without resetting or updating BL_0 using k_4 (i.e., the k that applies specifically to Timeline 1). OC behavior was detected in H_1 for most acquisition techniques initiating at 110 days (e.g., following the schema in Fig. 2, OC behavior was detected at 119 days corresponding to the second consecutive Cusum trigger at n. The reset takes place at $n-1$ corresponding to 110 days, which was the first Cusum trigger for this OC event) with these exceptions: W/Ag, 31 kV at 18 days; W/Ag, 32 kV at 10 days; and W/Rh, 27 kV at 91 days. In contrast, when using k_5 (i.e., the k that applies to Timeline 2), OC behavior was detected in H_1 only for the W/Ag, 27 and 28 kV techniques initiating at 110 days. Figure 4 (top-left) shows the DI monitoring of H_1 (W/Ag at 27 kV) for $h = 4$ cm (open dots) and

$h = 5$ cm (solid dots) using both $k_4 = 0.014$ and $k_5 = 0.021$, respectively, without resetting or updating BL_0. The downward arrows show deviations beyond the tolerance. The k_4 monitoring (open dots) indicates continuous OC behavior initiating at 110 days. In comparison when using k_5, OC behavior was detected once for four consecutive time points between 110 and 147 days (meeting the two-consecutive event criterion) and deviated past the tolerance randomly at four time points not meeting the two-consecutive event criterion at 216, 259, 357, and 427 days. The contrast between the 4 and 5 cm samples shows that at least part of the acquisition space can be OC while other parts are IC over extended periods. The corresponding DI plots with BL updating–resetting (i.e., adjusting m_0 and k_4, bringing BL_0 forward and resetting the Cusum argument to zero) are also shown in Fig. 4 (top-right), indicating three Cusum resets

TABLE II. Calibration variation due to 50/50 composition LRE shifts. This table illustrates the LRE shift (absolute value) required to induce a 4 PG shift for the ideal 50 PG value in the calibration for a specified height (h_s) and the associated deviation this shift translates to for other LRE samples for heights ranging from 3 to 7 cm. The italicized row shows that a 0.0383 LRE shift at 5 cm translates to a: (i) 4 PG shift at 5 cm; (ii) 5.71 PG shift at 3 cm; and (iii) 3.27 PG shift at 7 cm.

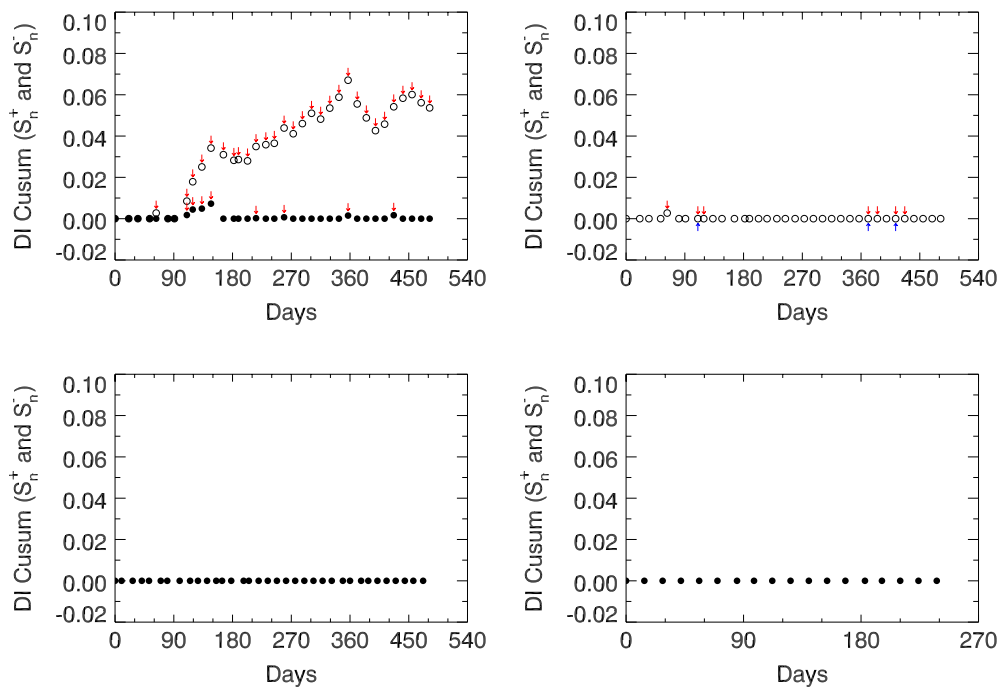| h_s (cm) | LRE shift at h_s giving a 4 PG change | Deviation from 50 PG at specified and nonspecified heights | | | | |
		3 cm	4 cm	5 cm	6 cm	7 cm
3	0.0268	4.00	3.26	2.80	2.45	2.29
4	0.0329	4.91	4.00	3.43	3.01	2.81
5	*0.0383*	*5.71*	*4.66*	*4.00*	*3.50*	*3.27*
6	0.0437	6.52	5.32	4.57	4.00	3.73
7	0.0469	6.99	5.70	4.90	4.29	4.00

FIG. 4. Decision interval cumulative sum examples for Timeline 1. The top-left plot shows H_1 monitoring (W/Ag at 27 kV) using k_4 (open dots) and k_5 (solid dots) without resetting or updating BL_0. The arrows (downward) indicate either a downward or upward drift exceeding the ±4 PG tolerance. The corresponding plots with BL updating and using the k (i.e., k_4) for 4 cm are shown in the top-right. Arrows pointing upward mark BL_0 updates and resets. The monitoring of H_2 is shown on the bottom-left (W/Ag at 27 kV) and for H_3 on the bottom-right (Mo/Mo at 29 kV), both using k_5. Note, the bottom plots show in control behavior over the entire time intervals. In these plots, S_n^+ and S_n^- are shown simultaneously noting that S_n^- was identically zero over the entire intervals.

were required for H_1 over this interval at 110, 371, and 413 days. Due to the similarity in the DI behavior and overlapping points, only one k_4 plot is shown. In this plot (top-right), the red downward arrows show calibration deviation past the tolerance and the blue upward arrows mark DI resets.

Due to similarities, plots for W/Ag at 28 kV for H_1 are not shown. In contrast with H_1, the H_2 and H_3 units were IC for all acquisition techniques for $h > 4$ cm (i.e., for all Timeline 1 and Timeline 2 images) as shown in the bottom plots in Fig. 4. To illustrate the influence of serial BL updating on k, we compare the adjusted LUTs at two time points in Fig. 5. This shows the LUTs for H_1 (W/Rh at 27 kV) at time zero and at 110 days (i.e., time of the update for 4 cm to monitor the 5 cm series). The curves diverge as h increases. In summary, adjusting k for $h \neq 4$ cm to monitor Timeline 1 images ($h = 4$ cm) establishes virtual timelines for a given height without performing additional imaging.

3.B. Serial calibration accuracy

The calibration accuracy evaluation with BL updating (within a given unit) due to OC behavior is not possible using Timeline 2 images at current. The required updating for H_1 with $h = 5$ cm was detected prior to the initiation of its Timeline 2 acquisitions, and the other units were IC over the interval. We used two alternative examples to show the validity of the updating approach. First to illustrate the merits of updating, we used Timeline 1 images from H_1 and constructed an artificial problem by not resetting the Cusum for the first two OC events. As indicated in Fig. 4 (top-right), OC behavior

was detected at 119 days (first reset event detected) and at 385 days (second reset event detected). We show the Cusum trajectory from the first event through 357 days (the time point before the second event) without resetting and adjusting in Fig. 6. Figure 6 (top) shows the portion of the DI for $h = 4$ cm (asterisks) fitted with linear regression (solid) distilled from Fig. 4 (H_1: W/Ag at 27 kV) corresponding to the OC interval. When the shift from the reference is approximated as a constant (not a function of time) at a given time point, the Cusum
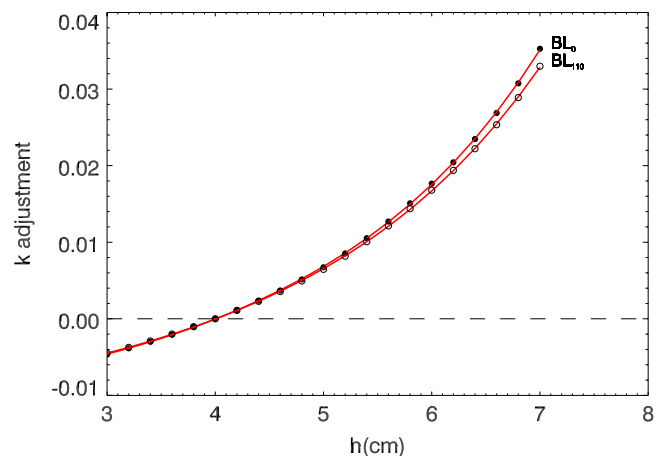

FIG. 5. Chart constant updating comparison. This shows the adjustments that are added to k_4 for arbitrary h used before (closed dots for BL_0) and after (open dots for BL_{110}) the DI reset at 110 days for W/AG at 27 kV. The curves diverge as the height (h) increases illustrating the necessity of the adjustment. The actual values for $k_4 = 0.0141$ (baseline) and $k_4 = 0.0138$ (110 days) were used to determine k_h ($h \neq 4$ cm) at the respective time point.
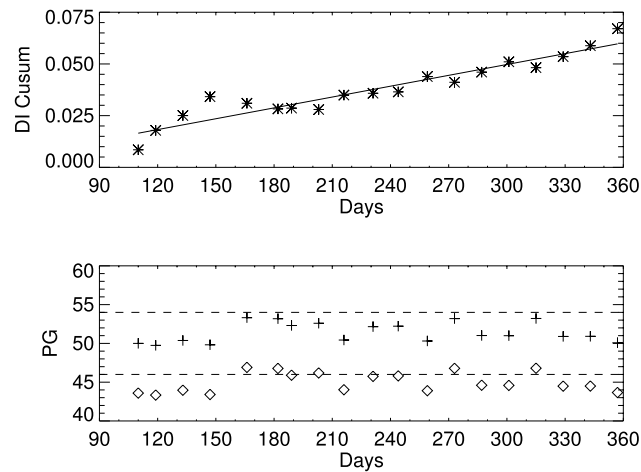
FIG. 6. DI cumulative sum monitoring and updating examples using Timeline 1. The top plot shows the DI (i.e., S_n^+) monitoring of H_1 for Timeline 1 (W/Ag at 27 kV) over a segment that exhibited out of control behavior (asterisks) starting at 110 days (top-plot) through 357 days. The fitted regression line (solid) is well approximated as linear (linear correlation ≈ 0.95) and indicates the shift was a constant jump detected at 119 days. The bottom plot shows the calibration accuracy using both the BL_0 datasets to calibrate the respective 50/50 compositions without updating (diamonds) and with updating (crosses) over the same interval. The dashed horizontal lines mark the upper and lower tolerances.

will exhibit a linear form afterward. The linear agreement of the fitted line (linear correlation = 0.95) with the observed sustained drift indicates the OC behavior was induced by a constant (approximately) jump in m_n detected at about 119 days (i.e., two consecutive points beyond the tolerance detected at n corresponding with 119 days). We updated BL_0 at 110 days and calibrated m_n for all n up to 357 days. The bottom plot in Fig. 6 shows the corresponding updated and nonupdated calibration accuracies for the same time period. The updated calibration values are within tolerance (dashed lines), whereas the majority (74%) of the calibration values derived from BL_0 are below the lower tolerance. Second, Fig. 7 shows the virtual monitoring of 3 cm 50/50 composition phantom images for H_1 (W/Ag at 27 kV), which shows OC behavior over most of the interval. The triangles on the x-axis mark the time points (days) of irregularly acquired 3 cm 50/50 composition phantom image acquisitions. These images were taken at irregular time points for other purposes. The system was IC for the first time point and OC for the two later time points. The calibration accuracy is shown in Table III, which indicates improved accuracy due to updating. The within-unit calibration accuracies for Timeline 2 are summarized in Table IV for all three units. Because the accuracies were similar, samples corresponding to every other kV are provided in Table IV, except where noted. The calibration results are presented as serial averages for the respective compositions and are all within tolerance.

The continual BL translation and cross-calibration accuracy summaries for the H_1 and H_2 units are provided in Table V. For either unit, the accuracies from cross-calibration derived from the other unit's Timeline 1 images are similar to those obtained from the within-unit calibration using continual BL translation. Because the findings across these two units for all
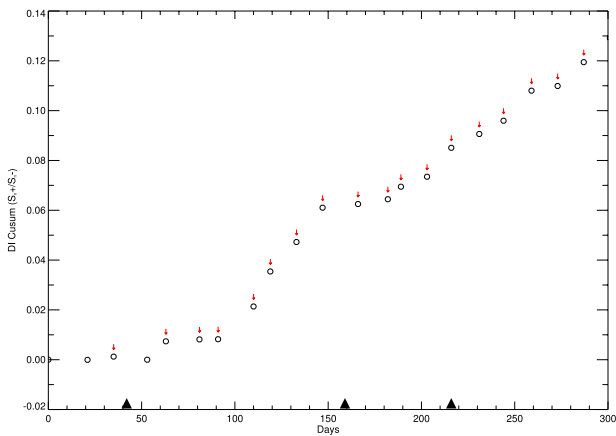

FIG. 7. DI cumulative sum monitoring of a 3 cm 50/50 composition. This shows a virtual 3 cm DI timeline for H_1 (W/Ag at 27 kV) derived from modifying the chart constant. The triangles on the x-axis mark the time points of the irregularly sampled LREs for a 3 cm 50/50 composition. The downward arrows indicate DI triggers. Thus, the two samples to the right were acquired when the system was out of control. Calibration results for these three samples are shown in Table III.

compositions were similar, we show without loss of generality, (i) the 20/80 and 40/60 compositions (i.e., less glandular content) for H_1 in Fig. 8 and (ii) the 60/40 and 80/20 compositions (greater glandular content) for H_2 in Fig. 9 (W/Ag at 27 kV). These plots illustrate (i) within-unit calibration for continual BL translation (diamonds), (ii) cross-calibration (plus signs), and (iii) calibration obtained by switching the respective BL_0 datasets (asterisks). In contrast, the accuracy obtained from switching the BL datasets (asterisks) is for the most part beyond tolerance. The continual BL translation findings for H_3 are summarized in Table VI (example plots not shown), similarly demonstrating the translation principle. Because the accuracies were similar, samples corresponding to every other kV are provided in Table VI. These findings indicate that the special case BL translation and the more general technique of cross-calibration as developed in Sec. 2.G. are valid approximations.

4. DISCUSSION

This paper introduced several novel concepts advancing both Cusum and calibration theory. The relative shift monitoring and LUT generalization represent key advancements in both areas. We modified the Cusum to detect a relative shift from the standard. Other researchers have modified the Cusum argument as well by normalizing it to the estimated deviation[40,41] but for different purposes. Our modification

TABLE III. 50/50 3 cm composition measured at irregular time points from H_1 calibrated without updating and with updating if *required.

Day	PG without update	PG with update
42	46.2	46.2
159	45.8	*51.7
216	44.5	*50.4

TABLE IV. Timeline 2 calibration accuracy. This table gives the within-unit calibration accuracy of Timeline 2 for the H_1, H_2, and H_3 units. The calibrated values for a given composition are presented as the mean values taken over the serial samples. The respective standard deviation of the mean distribution is provided parenthetically. The number of samples included for each composition for each unit is provided: H_1: $n_1 = 16$; H_2: $n_2 = 15$; and H_3: $n_3 = 16$. Serial updating was not required for most acquisition techniques except for those indicated[a]. Because the accuracies were similar across kV, every other kV sample is cited to reduce the presentation length with exception; all samples requiring updating are cited and marked.

FFDM unit	Target/filter	kV	20/80 composition	40/60 composition	60/40 composition	80/20 composition
H_1	W/Rh	26	19.5 (0.7)	40.1 (0.8)	60.2 (0.9)	79.4 (0.9)
		28	18.5 (0.5)	39.2 (0.6)	59.4 (0.6)	78.8 (0.5)
		30	18.2 (0.6)	38.5 (0.5)	58.1 (0.5)	77.0 (0.5)
	W/Ag	27[a]	22.1 (0.6)	42.6 (0.5)	62.9 (0.6)	82.0 (0.5)
		28[a]	22.4 (0.6)	42.8 (0.6)	62.9 (0.6)	81.7 (0.5)
		30	17.3 (0.6)	37.7 (0.6)	57.6 (0.6)	76.6 (0.6)
		32	20.6 (0.7)	40.5 (0.7)	60.0 (0.7)	78.8 (0.6)
H_2	W/Rh	26	18.7 (0.6)	39.6 (0.9)	60.3 (1.1)	79.9 (0.9)
		28	18.6 (0.9)	40.3 (0.5)	61.3 (0.7)	80.4 (1.7)
		30	19.0 (0.6)	40.2 (0.7)	61.1 (0.8)	81.0 (0.9)
	W/Ag	28	18.6 (0.7)	39.6 (0.7)	60.8 (0.8)	80.5 (0.8)
		30	18.9 (0.7)	39.3 (0.7)	60.0 (0.5)	79.3 (0.7)
		32	21.0 (0.7)	40.9 (0.5)	60.9 (0.5)	79.7 (0.5)
H_3	Mo/Mo	26	21.6 (0.5)	42.1 (0.4)	62.4 (0.5)	80.5 (0.8)
		28	21.1 (0.4)	41.6 (0.3)	62.5 (0.6)	81.2 (0.5)
		30	21.9 (0.4)	42.1 (0.3)	62.7 (0.5)	81.0 (0.5)
	Mo/Rh	26	21.1 (0.5)	41.4 (0.4)	62.2 (0.6)	80.9 (0.5)
		28	20.5 (0.5)	40.9 (0.6)	62.0 (0.5)	81.3 (0.5)
		30	21.4 (0.6)	41.1 (0.6)	61.7 (0.5)	80.5 (0.5)
		32	21.2 (0.5)	41.2 (0.6)	62.1 (0.6)	81.1 (0.6)

[a]BL_0 was updated to BL_{110}; Timeline 2 for H_1 was initiated approximately at day 273.

was for a secondary application of maintaining prospective calibration accuracy. The relative shift monitoring provided a foundation for developing a method of determining a suitable chart constant with numerical analyses tied to the calibration accuracy. We showed that calibration accuracy is a function of the acquisition technique as well as the compressed breast thickness (i.e., phantom height in this paper) with respect to a given LRE standard. The calibration accuracy is generally more sensitive to LRE shifts for smaller compressed breast thicknesses and for higher energy beams. This is expected because the separation between the adipose and glandular calibration curves decreases as thickness decreases and the energy of the beam increases. The chart constant was adjusted to achieve uniform serial calibration accuracy across the acquisition techniques and heights. Effectively in the prospective application, each patient will have their own monitoring timeline based on the mammography unit's compressed breast thickness reading.

A system to monitor and maintain serial calibration accuracy was presented and evaluated with three FFDM units. Baseline translation requires one serial image for each acquisition technique per FFDM unit. These images are from the Timeline 1 serial monitoring dataset. Therefore, additional imaging is not required for BL translation. Generally, the serial calibration accuracy was within the ±4 PG tolerance as demonstrated when calibrating Timeline 2 images for all units. This

BL translation principle was validated with the H_1 unit using a segment of its Timeline 1 images, spanning approximately 210 days. In this example, BL translation corrected the calibrated timeline segment by moving the respective accuracy into the IC region. To further demonstrate the BL translation, we set up an artificial application. This application included updating the BL continually at each time point, n, and calibrating the Timeline 2 images at each respective n, which gave similar calibration accuracies as derived from the standard application for a given unit indicating equivalency between the many BL datasets. The BL translation is a special case of the more general cross-calibration principle that shows how to convert the BL dataset of one unit to the other unit's (unit 2) BL at arbitrary n by using the Timeline 1 images acquired at n with unit 2. The cross-calibration application was validated by showing its accuracy was similar to that of the continual BL updating (within-unit) and was within the ±4 PG tolerance. As a counter example, we switched the BL calibration datasets (Cal-switch) for H_1 and H_2 and applied the calibration without adjustments. As expected, the accuracy was outside of the tolerance region for the majority of samples because in general, the BL_n from unit 1 is not interchangeable with the BL_n from unit 2 or vice versa.

There are several aspects of this work that warrant further comment. An automated breast density measure suitable for

TABLE V. Baseline calibration dataset translation and cross-calibration evaluation. For each unit (H_1, and H_2), the baseline dataset BL_0 (i.e., time zero) was translated forward in time and updated to BL_n with Timeline 1 images on a continual biweekly basis from day 0 to approximately 482 days. BL_n was then used to calibrate the corresponding Timeline 2 images at each n (within-unit). For the cross-calibration experiment: (1) the BL_0 dataset for H_1 was adjusted with Timeline 1 images acquired with H_2 and then used to calibrate the Timeline 2 images acquired with H_2 (H_2 cross-cal); (2) the BL_0 dataset for H_2 was adjusted with Timeline 1 images acquired with H_1 and used to calibrate the Timeline 2 images acquired with H_1 (H_1 cross-cal); and (3) for comparison, the BL_0 from H_1 was used to calibrate Timeline 2 images acquired with H_2 and vice versa (Cal-switch). Serial averages are provided in this table. The respective standard deviation of the mean distribution is provided parenthetically. The number of serial samples for each composition for each unit is provided: H_1: $n_1 = 16$; and H_2: $n_2 = 15$.

FFDM unit	Target/filter	kV	20/80 composition Within-unit	Cross-cal	Cal-switch	40/60 composition Within-unit	Cross-cal	Cal-switch	60/40 composition Within-unit	Cross-cal	Cal-switch	80/20 composition Within-unit	Cross-cal	Cal-switch
H_1	W/Rh	26	21.6 (2.1)	20.6 (2.1)	22.3 (0.7)	42.2 (1.9)	41.4 (1.9)	43.2 (0.8)	62.3 (2.1)	61.8 (2.2)	63.5 (0.9)	81.5 (2.4)	81.2 (2.4)	82.9 (1.0)
		27	21.9 (2.0)	22.3 (2.0)	22.8 (0.6)	42.2 (1.8)	42.6 (1.9)	43.2 (0.6)	61.9 (2.0)	62.4 (2.0)	62.9 (0.7)	80.7 (2.0)	81.2 (2.0)	81.7 (0.6)
		28	20.7 (1.7)	21.3 (1.8)	22.6 (0.5)	41.4 (1.7)	42.7 (1.7)	44.0 (0.6)	61.6 (1.9)	63.5 (1.9)	64.8[a] (0.6)	81.0 (1.8)	83.5 (1.9)	84.8[a] (0.5)
		29	21.3 (1.6)	21.9 (1.6)	23.8 (0.6)	41.8 (1.3)	42.9 (1.3)	44.9[a] (0.6)	61.6 (1.5)	63.3 (1.5)	65.2[a] (0.6)	80.7 (1.7)	82.9 (1.8)	84.8[a] (0.7)
		30	20.6 (1.5)	20.4 (1.6)	22.3 (0.6)	40.9 (1.3)	41.9 (1.3)	43.8 (0.6)	60.5 (1.5)	62.7 (1.5)	64.6[a] (0.5)	79.4 (1.6)	82.6 (1.7)	84.5[a] (0.6)
	W/Ag	27	20.2 (1.5)	20.6 (1.6)	15.5[a] (0.6)	40.7 (1.5)	41.7 (1.6)	36.6 (0.5)	60.9 (1.7)	62.6 (1.8)	57.5 (0.6)	80.1 (1.6)	82.3 (1.7)	77.2 (0.5)
		28	20.5 (1.5)	19.5 (1.6)	13.8[a] (0.6)	41.0 (1.3)	41.0 (1.4)	35.3[a] (0.6)	61.0 (1.5)	62.1 (1.6)	56.4 (0.6)	79.9 (1.6)	82.0 (1.7)	76.3 (0.6)
		29	20.3 (1.5)	20.3 (1.5)	14.5[a] (0.6)	40.5 (1.3)	40.9 (1.3)	35.1[a] (0.7)	60.3 (1.4)	61.2 (1.5)	55.4[a] (0.6)	79.1 (1.6)	80.5 (1.6)	74.7[a] (0.6)
		30	19.5 (1.4)	19.4 (1.4)	13.4[a] (0.6)	39.8 (1.2)	40.2 (1.3)	34.2[a] (0.6)	59.8 (1.4)	60.6 (1.4)	54.6[a] (0.6)	78.7 (1.5)	80.1 (1.6)	74.0[a] (0.7)
		31	19.6 (1.4)	20.4 (1.4)	14.1[a] (0.7)	39.4 (1.1)	40.5 (1.2)	34.2[a] (0.7)	58.9 (1.3)	60.2 (1.4)	54.0[a] (0.7)	77.4 (1.5)	79.0 (1.5)	72.7[a] (0.6)
		32	19.9 (1.2)	19.5 (1.2)	14.4[a] (0.7)	39.7 (1.2)	39.5 (1.2)	34.4[a] (0.7)	59.3 (1.4)	59.3 (1.4)	54.2[a] (0.7)	78.0 (1.3)	78.2 (1.3)	73.1[a] (0.6)
H_2	W/Rh	26	19.3 (1.2)	20.4 (1.2)	15.9[a] (0.6)	40.2 (1.3)	41.0 (1.3)	36.6 (0.9)	60.9 (1.6)	61.5 (1.5)	57.1 (1.0)	80.5 (1.4)	80.9 (1.4)	76.4 (0.9)
		27	20.1 (1.5)	19.7 (1.5)	14.7[a] (1.0)	40.5 (1.6)	40.1 (1.6)	35.1[a] (0.6)	60.5 (1.9)	60.0 (1.9)	55.0[a] (1.0)	78.6 (2.1)	78.0 (2.1)	73.1[a] (2.1)
		28	19.2 (1.5)	18.6 (1.5)	14.7[a] (0.8)	40.9 (1.3)	39.6 (1.2)	35.6[a] (0.4)	61.9 (1.4)	60.0 (1.4)	56.0 (0.6)	81.0 (2.0)	78.5 (1.9)	74.5[a] (1.6)
		29	20.1 (1.5)	19.6 (1.5)	15.0[a] (0.9)	40.9 (0.9)	39.8 (0.9)	35.2[a] (0.5)	61.4 (1.2)	59.8 (1.2)	55.2[a] (0.6)	81.0 (1.9)	78.9 (1.9)	74.3[a] (1.2)
		30	19.0 (1.1)	19.3 (1.0)	15.1[a] (0.6)	40.2 (1.2)	39.4 (1.1)	35.1[a] (0.6)	61.2 (1.4)	59.1 (1.4)	54.9[a] (0.8)	81.1 (1.4)	77.9 (1.3)	73.6[a] (0.9)
	W/Ag	27	18.9 (1.2)	18.6 (1.1)	26.4[a] (0.8)	39.7 (1.1)	38.7 (1.0)	46.6[a] (0.7)	60.7 (1.2)	59.1 (1.1)	66.9[a] (0.7)	80.2 (1.2)	78.1 (1.2)	85.9[a] (0.7)
		28	17.9 (1.1)	19.0 (1.1)	26.9[a] (0.6)	38.9 (1.3)	39.0 (1.2)	46.9[a] (0.6)	60.1 (1.5)	59.1 (1.5)	67.0[a] (0.8)	79.8 (1.4)	77.8 (1.3)	85.6[a] (0.8)
		29	19.1 (1.0)	19.2 (1.0)	22.3 (0.6)	39.3 (1.0)	38.9 (1.0)	42.0 (0.7)	59.7 (1.0)	58.8 (1.0)	61.9 (0.6)	78.8 (1.0)	77.5 (1.0)	80.6 (0.6)
		30	18.3 (1.1)	18.4 (1.1)	22.7 (0.6)	38.7 (1.1)	38.4 (1.1)	42.7 (0.7)	59.4 (1.2)	58.5 (1.1)	62.8 (0.5)	78.7 (1.2)	77.4 (1.2)	81.7 (0.6)
		31	19.3 (1.1)	18.5 (1.1)	27.1[a] (0.8)	39.1 (0.9)	38.1 (0.9)	46.7[a] (0.6)	59.1 (1.1)	57.8 (1.1)	66.4[a] (0.7)	77.8 (1.1)	76.2 (1.0)	84.8[a] (0.6)
		32	18.3 (1.1)	18.7 (1.1)	27.2[a] (0.7)	38.2 (1.0)	38.4 (1.0)	46.9[a] (0.5)	58.2 (1.2)	58.2 (1.1)	66.7[a] (0.5)	77.0 (1.1)	76.9 (1.1)	85.4[a] (0.5)

Note: Within-unit: Calibration curves were derived with continual biweekly BL translation using the Timeline 1 images from the same FFDM unit; Cross-cal: BL_0 calibration curves were switched, adjusted serially with the other unit's Timeline 1 images, and used on the unit's Timeline 2 images; Cal-switch: BL_0 calibration datasets were switched at time zero and used on the other unit to calibrate all Timeline 2 images.

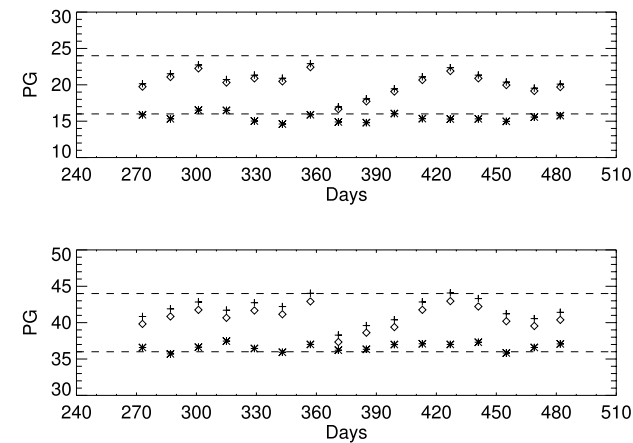[a]Calibrated quantities that were outside of the ± 4 PG tolerance.

FIG. 8. Timeline 2 baseline translation and cross-calibration examples from H_1. Calibrated Timeline 2 images acquired with H_1 (W/Ag at 27 kV) are shown: 20/80 (top) and 40/60 (bottom) compositions. In these plots, the calibration was based on (i) the continuous baseline translation within-unit (diamonds) (ii) cross-calibration using BL_0 calibration data acquired from H_2 with continual updating using Timeline 1 images acquired from H_1 (plus signs), and (iii) switching the calibration data (asterisks) by calibrating Timeline 2 images acquired with H_1 with the BL_0 data taken with H_2.
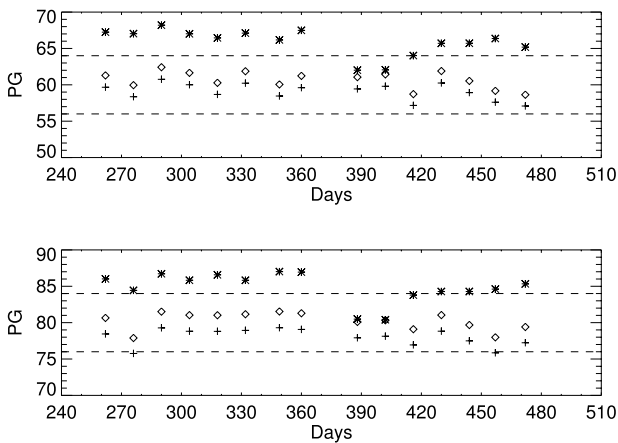
FIG. 9. Timeline 2 baseline translation and cross-calibration examples from H_2. Calibrated Timeline 2 images acquired with H_2 (W/Ag at 27 kV) are shown: 60/40 (top) and 80/20 (bottom) compositions. In these plots, the calibration was based on (i) the continuous baseline translation within-unit (diamonds) (ii) cross-calibration using BL_0 calibration data acquired from H_1 with continual updating using Timeline 1 images acquired from H_2 (plus signs), and (iii) switching the calibration data (asterisks) by calibrating Timeline 2 images acquired with H_2 with the BL_0 data taken with H_1.

the clinical environment should have a high degree of reproducibility both within and across imaging platforms without undue modification. The monitoring and updating techniques presented in this paper address the serial reproducibility attribute. To achieve uniform serial calibration accuracy (reproducibility), we showed that both height (i.e., compressed breast thickness) and the acquisition technique must be taken into account. Similarly, due to target/filter differences, the adipose and glandular calibration curves for H_3 are separated further than those for H_1 and H_2 for a given height, indicating that accuracy of H_3 is less affected by serial LRE shifts. If not taken into account, these differences could affect analyses that merge images from different units. We noted more variation in H_1. This is the older unit, installed in July 2006. The H_1 detector was replaced in November 2010 (about two years prior to its first OC event) and March 2014 (about 9 months after its second OC event). To date, H_1 passed all of its weekly flat

field evaluations. This system has always passed the Mammography Quality Standard Act evaluation performed annually in September and the physicist's annual inspection each October. The similar H_2 unit was installed in July 2009 and passed the same weekly and annual evaluations. H_2 has not had its detector replaced. The H_3 unit was installed in December 2012 and performed as H_2 under the same evaluations and has not had its detector replaced. In previous work[17] with different FFDM technologies, we noted target/filter dependent OC behavior prior to the x-ray tube failing, which is in contrast with the OC behavior noted in this paper for H_1. At this time, it is not clear why H_1 exhibited different behavior than the other two units. However, we speculate that variation across the detectors may account for the noted differences. The cross-calibration application addressed interunit generalization. Thus, it should be possible to establish complete sets of BL calibration curves without acquiring full calibration datasets for similar FFDM

TABLE VI. Baseline calibration dataset with continual translation for the H_3 unit. For the H_3 unit, the baseline dataset BL_0 (i.e., time zero) was translated forward in time and updated with the Timeline 1 images (from the same unit) biweekly giving BL_n from day 0 to 238 days. BL_n was then used to calibrate the corresponding Timeline 2 phantoms at each time point $= n$ (within-unit). Serial averages are provided in this table. The respective standard deviation of the mean distribution is provided parenthetically. Because the accuracies were similar across kV, every other kV sample is shown. There are 16 serial samples per composition.

FFDM unit	Target/filter	kV	20/80 composition	40/60 composition	60/40 composition	80/20 composition
H_3	Mo/Mo	26	20.7 (0.8)	41.1 (0.8)	61.4 (0.8)	79.5 (1.0)
		28	19.8 (0.9)	40.3 (0.8)	61.2 (0.8)	79.9 (0.7)
		30	19.8 (0.9)	40.0 (0.9)	60.5 (0.8)	78.9 (0.8)
	Mo/Rh	26	19.6 (0.8)	39.9 (0.8)	60.6 (0.9)	79.3 (0.8)
		28	19.1 (0.8)	39.5 (1.0)	60.6 (0.8)	79.9 (0.8)
		30	19.9 (1.4)	39.6 (1.4)	60.2 (1.2)	79.0 (1.4)
		32	19.8 (1.0)	39.8 (1.0)	60.7 (0.9)	79.6 (0.9)

units. We also demonstrated that the calibration approach developed with indirect x-ray conversion FFDM applies to direct x-ray conversion units as well with refinements, also addressing intertechnology generalization.

There are also several qualifications with our findings. The reported accuracies are best considered ideal because phantoms with precise heights were used in the evaluation. The BL updating illustrations for the most part pertained to phantom heights at 4 and 5 cm. However, the limited number of 3 cm samples also indicates the approach is valid over the height range. Uncertainty caused by variation in compressed breast thickness did not impact our analysis as would be the case when calibrating mammograms. We are developing a compressed breast thickness correction to compensate for uncertainty introduced by the compression paddle for each unit based on our related work.[3] Generally, the compressed breast thickness correction is a function of the pixel coordinates defining a surface. The method for incorporating a correction surface into our serial monitoring is under development. As noted previously,[13] the x-ray attenuation of the adipose BTE material does not match that of the adipose breast tissue precisely. This attenuation artifact was not in effect in this study, as would be encountered when calibrating mammograms using BTE phantoms. Although we have addressed this attenuation artifact previously,[13] we believe more analysis is required to find an approximate solution. Both the translation and cross-calibration techniques were derived by assuming differences in the BL data acquired at different times differed by a constant offset observable in the serial LRE measurements from Timeline 1 and not where drift is a function of time. As observed previously, DI monitoring exhibited nonlinear behavior (i.e., drift was a function of time) prior to x-ray tube failure when analyzing data from another FFDM technology.[17] This artifact is mitigated by our two-consecutive event criterion definition for OC behavior intervention. As noted,[17] a constant LRE shift gives rise to a linear DI trajectory. To detect this form of drift, the OC behavior can be evaluated for linearity as part of the monitoring process.

5. CONCLUSION

We presented a methodology to maintain serial calibration accuracy using phantom images acquired with direct x-ray conversion FFDM. Monitoring techniques developed earlier were extended by generalizing the DI chart constant with LUTs, permitting the surveillance of many processes simultaneously without requiring additional imaging. The approach relies on acquiring one phantom image per acquisition technique for each FFDM unit biweekly, translating into approximately 20 minutes effort per unit. We evaluated a novel method of updating calibration datasets using the images acquired for the serial monitoring (i.e., with no additional effort). This updating translates the current BL dataset forward in time with a minimal amount of imaging effort. Baseline translation is a special case of a more general cross-calibration application, which may be useful for applying calibration at facilities with similar units, eliminating the necessity of acquiring full

BL datasets. Further evaluation of the serial monitoring and BL update/translation theory across other sites and imaging technologies is required to substantiate its generality. Tomosynthesis (TS) is an emerging imaging technology[42] that may supplant or at least exist in parallel with conventional two-dimensional (2D) FFDM. As raw TS data is comprised of multiple 2D projections,[43] we would expect, by hypothesis, our calibration methodology to translate. Most importantly, our techniques will require validation using patient images with breast cancer status as the endpoint comparison, which is the next planned investigation.

ACKNOWLEDGMENT

APPENDIX: CHART CONSTANT ESTIMATES

The 50/50 reference LREs for 4 cm are established from the BL_0 dataset using a linear combination with equal weights = 1/2 because these are not included in the BL datasets. We include a height dependency in m_n [i.e., $m_n(h)$]. To estimate k numerically for $h = 4$ cm, we use Eq. (1) and shift $m_n(4)$ incrementally giving $m_n(4) + \alpha_j$, where j is the increment index and α_j is the increment at j. We estimate numerically $CAL[m_n(4) + \alpha_j]$ that gives (50 ± 4) PG. Due to symmetry, either 46 PG or 54 PG can be used equivalently for the k solution. The numerical solution for specific $\alpha_j = d_4$ is then used to express $k_4 = d_4/|m_n(4)|$ for Timeline 1. The same approach was used to determine k for any height (i.e., $h \neq 4$) that is then used with the Timeline 1 monitoring, creating virtual timelines corresponding to any h as required. Using the same approach for $h \neq 4$ gives d_h such that $CAL[m_n(h) + d_h] = PG_{ideal} \pm 4$ PG for all h. The modified k used with Timeline 1 (i.e., the 4 cm timeline) is then given by $k_h = d_h/|m_n(4)|$ producing a virtual timeline for monitoring samples (images) for a given height referenced to the preset tolerance. We present LUTs for the k_h determination as well. When applying this technique to virtual timelines (i.e., $h \neq 4$ cm), the required m_n is also estimated from the appropriate calibration curves as a linear combination, if not imaged or measured explicitly.

a)Authors to whom correspondence should be addressed. Electronic addresses: erin.fowler@moffitt.org and john.heine@moffitt.org

[1]B. M. Keller, D. L. Nathan, Y. Wang, Y. Zheng, J. C. Gee, E. F. Conant, and D. Kontos, "Estimation of breast percent density in raw and processed full field digital mammography images via adaptive fuzzy c-means clustering and support vector machine segmentation," Med. Phys. **39**, 4903–4917 (2012).

[2]M. J. Yaffe, "Mammographic density Measurement of mammographic density," Breast Cancer Res. **10**, 209–218 (2008).

[3]J. J. Heine, K. Cao, and J. A. Thomas, "Effective radiation attenuation calibration for breast density: Compression thickness influences and correction," Biomed. Eng. Online **9**, 73–98 (2010).

[4]J. J. Heine and J. A. Thomas, "Effective x-ray attenuation coefficient measurements from two full field digital mammography systems for data calibration applications," Biomed. Eng. Online **7**, 13–24 (2008).

[5]R. Highnam and M. Brady, *Mammographic Image Analysis* (Kluwer Academic, Boston, MA, 1999).

[6] R. Highnam, X. Pan, R. Warren, M. Jeffreys, G. D. Smith, and M. Brady, "Breast composition measurements using retrospective standard mammogram form (SMF)," Phys. Med. Biol. **51**, 2695–2713 (2006).

[7] J. Kaufhold, J. A. Thomas, J. W. Eberhard, C. E. Galbo, and D. E. Trotter, "A calibration approach to glandular tissue composition estimation in digital mammography," Med. Phys. **29**, 1867–1880 (2002).

[8] S. Malkov, J. Wang, K. Kerlikowske, S. R. Cummings, and J. A. Shepherd, "Single x-ray absorptiometry method for the quantitative mammographic measure of fibroglandular tissue volume," Med. Phys. **36**, 5525–5536 (2009).

[9] O. Pawluczyk, B. J. Augustine, M. J. Yaffe, D. Rico, J. Yang, G. E. Mawdsley, and N. F. Boyd, "A volumetric method for estimation of breast density on digitized screen-film mammograms," Med. Phys. **30**, 352–364 (2003).

[10] S. van Engeland, P. R. Snoeren, H. Huisman, C. Boetes, and N. Karssemeijer, "Volumetric breast density estimation from full-field digital mammograms," IEEE Trans. Med. Imaging **25**, 273–282 (2006).

[11] S. Ciatto, D. Bernardi, M. Calabrese, M. Durando, M. A. Gentilini, G. Mariscotti, F. Monetti, E. Moriconi, B. Pesce, A. Roselli, C. Stevanin, M. Tapparelli, and N. Houssami, "A first evaluation of breast radiological density assessment by QUANTRA software as compared to visual classification," Breast **21**, 503–506 (2012).

[12] R. Highnam, N. Sauber, S. Destounis, J. Harvey, and D. Mcdonald, in *Presented at the 11th International Workshop of Digital Mammography, Philadelphia, PA* (2012) (unpublished).

[13] J. J. Heine, K. Cao, D. E. Rollison, G. Tiffenberg, and J. A. Thomas, "A quantitative description of the percentage of breast density measurement using full-field digital mammography," Acad. Radiol. **18**, 556–564 (2011).

[14] E. E. Fowler, B. Lu, and J. J. Heine, "A comparison of calibration data from full field digital mammography units for breast density measurements," Biomed. Eng. Online **12**, 114–135 (2013).

[15] J. J. Heine and M. Behera, "Effective x-ray attenuation measurements with full field digital mammography," Med. Phys. **33**, 4350–4366 (2006).

[16] D. M. Hawkins and D. H. Olwell, *Cumulative Sum Charts and Charting for Quality Improvement* (Springer-Verlag, New York, NY, 1997).

[17] J. J. Heine, K. Cao, and C. Beam, "Cumulative sum quality control for calibrated breast density measurements," Med. Phys. **36**, 5380–5390 (2009).

[18] E. S. Page, "Continuous inspection schemes," Biometrika **41**, 100–115 (1954).

[19] E. S. Page, "Cumulative sum schemes using gauging," Technometrics **4**, 97–109 (1962).

[20] E. S. Page, "Controlling the standard deviation by Cusums and warning lines," Technometrics **5**, 307–315 (1963).

[21] A. Bartlett and B. Parry, "Cusum analysis of trends in operative selection and conversion rates for laparoscopic cholecystectomy," Aust. N. Z. J. Surg. **71**, 453–456 (2001).

[22] F. Claerhout, J. P. Roovers, P. Lewi, J. Verguts, D. De Ridder, and J. Deprest, "Implementation of laparoscopic sacrocolpexy—A single centre's experience," Int. Urogynecol. J. Pelvic Floor Dysfunct. **20**, 1119–1125 (2009).

[23] M. R. de Leval, K. François, C. Bull, W. Brawn, and D. Spiegelhalter, "Analysis of a cluster of surgical failures: Application to a series of neonatal arterial switch operations," J. Thorac. Cardiovasc. Surg. **107**, 914–924 (1994).

[24] G. R. de Oliveira Filho, "The construction of learning curves for basic skills in anesthetic procedures: An application for the cumulative sum method," Anesth. Analgesia. **95**, 411–416 (2002).

[25] G. L. Grunkemeier, Y. X. Wu, and A. P. Furnary, "Cumulative sum techniques for assessing surgical results," Ann. Thorac. Surg. **76**, 663–667 (2003).

[26] D. M. Holzhey, J. Seeburger, M. Misfeld, M. A. Borger, and F. W. Mohr, "Learning minimally invasive mitral valve surgery: A cumulative sum sequential probability analysis of 3895 operations from a single high-volume center," Circulation **128**, 483–491 (2013).

[27] R. J. Novick, S. A. Fox, L. W. Stitt, B. B. Kiaii, S. A. Swinamer, R. Rayman, T. R. Wenske, and W. D. Boyd, "Assessing the learning curve in off-pump coronary artery surgery via CUSUM failure analysis," Ann. Thorac. Surg. **73**, S358–S362 (2002).

[28] P. Rozenberg, R. Porcher, L. J. Salomon, F. Boirot, C. Morin, and Y. Ville, "Comparison of the learning curves of digital examination and transabdominal sonography for the determination of fetal head position during labor," Ultrasound Obstet. Gynecol. **31**, 332–337 (2008).

[29] S. E. Smith, V. R. Tallentire, J. Spiller, S. M. Wood, and H. S. Cameron, "The educational value of using cumulative sum charts," Anaesthesia **67**, 734–740 (2012).

[30] P. P. Tekkis, V. W. Fazio, I. C. Lavery, F. H. Remzi, A. J. Senagore, J. S. Wu, S. A. Strong, J. D. Poloneicki, T. L. Hull, and J. M. Church, "Evaluation of the learning curve in ileal pouch-anal anastomosis surgery," Ann. Surg. **241**, 262–268 (2005).

[31] L. Sun, K. Wilder, and M. S. Mcpeek, "Enhanced pedigree error detection," Hum. Hered. **54**, 99–110 (2002).

[32] J. Sabria, C. Barcelo-Vidal, M. Arigita, J. M. Jimenez, B. Puerto, and A. Borrell, "The CUSUM test applied in prospective nuchal translucency quality review," Ultrasound Obstet. Gynecol. **37**, 582–587 (2011).

[33] B. J. Cowling, I. O. Wong, L. M. Ho, S. Riley, and G. M. Leung, "Methods for monitoring influenza surveillance data," Int. J. Epidemiol. **35**, 1314–1321 (2006).

[34] M. Hohle, M. Paul, and L. Held, "Statistical approaches to the monitoring and surveillance of infectious diseases for veterinary public health," Prev. Vet. Med. **91**, 2–10 (2009).

[35] J. Menotti, R. Porcher, P. Ribaud, C. Lacroix, V. Jolivet, S. Hamane, and F. Derouin, "Monitoring of nosocomial invasive aspergillosis and early evidence of an outbreak using cumulative sum tests (CUSUM)," Clin. Microbiol. Infect. **16**, 1368–1374 (2010).

[36] P. A. Rogerson and I. Yamada, "Approaches to syndromic surveillance when data consist of small regional counts," MMWR Morb Mortal Wkly Rep **53**(suppl), 79–85 (2004).

[37] P. A. Rogerson and I. Yamada, "Monitoring change in spatial patterns of disease: Comparing univariate and multivariate cumulative sum approaches," Stat. Med. **23**, 2195–2214 (2004).

[38] B. Barratt, R. Atkinson, H. Ross Anderson, S. Beevers, F. Kelly, I. Mudway, and P. Wilkinson, "Investigation into the use of the CUSUM technique in identifying changes in mean air pollution levels following introduction of a traffic management scheme," Atmos. Environ. **41**, 1784–1791 (2007).

[39] A. Schuh, J. A. Camelio, and W. H. Woodall, "Control charts for accident frequency: A motivation for real-time occupational safety monitoring," Int. J. Inj. Contr. Saf. Promot. **21**, 154–162 (2014).

[40] T. O. Lim, A. Soraya, L. M. Ding, and Z. Morad, "Assessing doctors' competence: Application of CUSUM technique in monitoring doctors' performance," Int. J. Qual. Health Care **14**, 251–258 (2002).

[41] N. M. Ung and L. Wee, "Cumulative sum method in nonzero fixed action level setup correction strategy: An application of statistical process control for targeted prostate radiotherapy," Med. Phys. **39**, 2746–2753 (2012).

[42] B. Ren, C. Ruth, T. Wu, Y. Zhang, A. Smith, L. Niklason, C. Williams, E. Ingal, B. Polischuk, and Z. Jing, presented at the Medical Imaging 2010: Physics of Medical Imaging, San Diego, CA (2010) (unpublished).

[43] A. Smith, "Fundamentals of breast tomosynthesis," in a series of research overviews on advanced technologies in women's healthcare, 2008.